Supplement:
Batch effects analysis for TCGA Thyroid cancer data sets

Rehan Akbani, Shiyun Ling, John N. Weinstein
Dept. of Bioinformatics and Computational Biology,
University of Texas MD Anderson Cancer Center – Genome Data Analysis Center

Supplemental Methods:

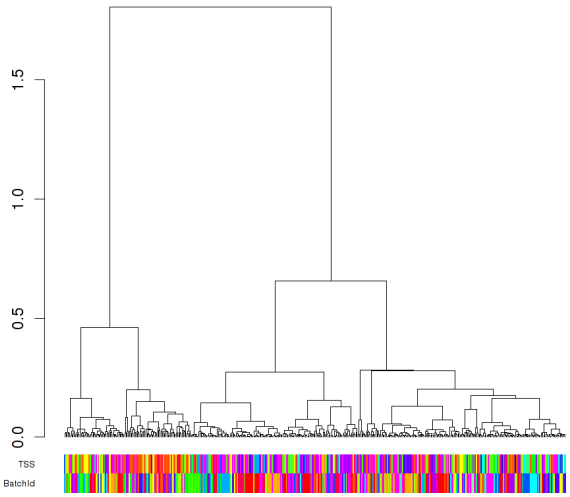
We used hierarchical clustering and Principal Components Analysis (PCA) to assess batch effects in the thyroid cancer data sets. Four different data sets were analyzed: miRNA sequencing (Illumina HiSeq), DNA methylation (Infinium HM450 microarray), mRNA sequencing (Illumina HiSeq), and protein expression (RPPA). All of the data sets were at TCGA level 3, since that's the level at which most of the analyses in the paper are based. We assessed batch effects with respect to two variables; batch ID and Tissue Source Site (TSS). Detailed results and batch effects analysis of other TCGA data sets can be found at: http://bioinformatics.mdanderson.org/tcgabatcheffects

For hierarchical clustering, we used the average linkage algorithm with 1 minus the Pearson correlation coefficient as the dissimilarity measure. We clustered the samples and then annotated them with colored bars at the bottom. Each color corresponded to a batch ID or a TSS. For PCA, we plotted the first four principal components, but only plots of two components are shown here. To make it easier to assess batch effects, we enhanced the traditional PCA plot with centroids. Points representing samples with the same batch ID (or TSS) were connected to the batch centroid by lines. The centroids were computed by taking the mean across all samples in the batch. That procedure produced a visual representation of the relationships among batch centroids in relation to the scatter within batches. The results for the four data sets follow.

miRNA (RNA-seq Illumina HiSeq)

Figures 1-3 show clustering and PCA plots for miRNA seq data. miRNAs with zero values were removed and the read counts were $\log_2$-transformed before generating the figures. The figures show no major batch effects and none of the batches or TSSs stand out from the rest.

Legends



BatchId

- 0087 (24)
- 115 (14)
- 126 (21)
- 149 (26)
- 171 (23)
- 176 (24)
- 182 (59)
- 191 (38)
- 197 (21)
- 205 (23)
- 224 (26)
- 230 (35)
- 237 (18)
- 250 (59)
- 253 (16)

TSS

- BJ – University of Pittsburgh (45)
- CE – ILSBio (8)
- DE – University of North Carolina (5)
- DJ – Memorial Sloan Kettering (82)
- DO – Medical College of Georgia (3)
- E3 – Roswell Park (7)
- E8 – Asterand (17)
- EL – MD Anderson (65)
- EM – University Health Network (70)
- ET – Johns Hopkins (53)
- FE – Ohio State University (16)
- FK – Johns Hopkins (6)
- FY – International Genomics Consortium (23)
- GE – ABS – IUPUI (1)
- H2 – Christiana Healthcare (6)
- IM – University of Miami (6)
- J8 – Mayo Clinic (10)
- KS – University Of Michigan (4)

Fig. 1. Hierarchical clustering for miRNA expression from miRNA-seq data

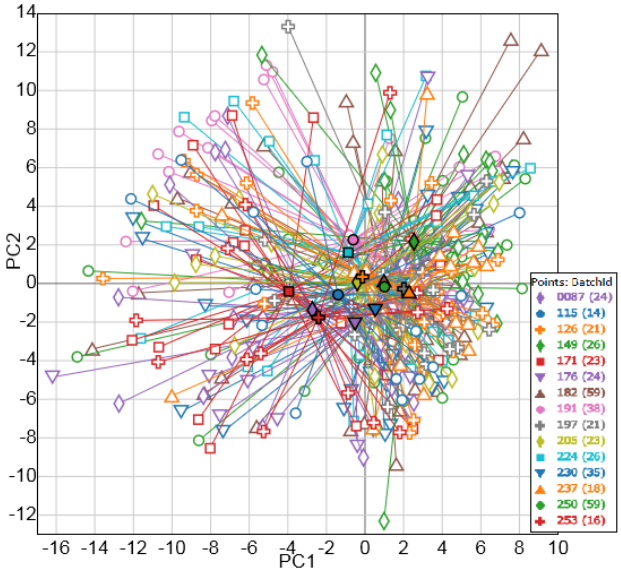
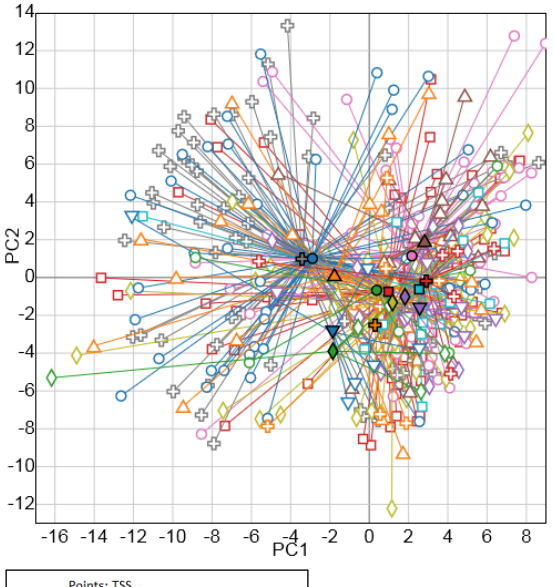


Fig. 2. PCA: First two principal components for miRNA expression from miRNA-seq data, with samples connected by centroids according to batch ID.



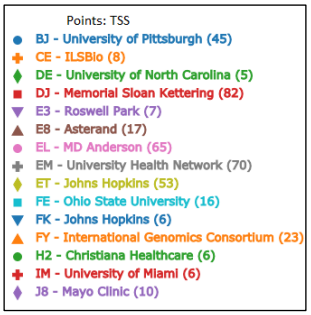
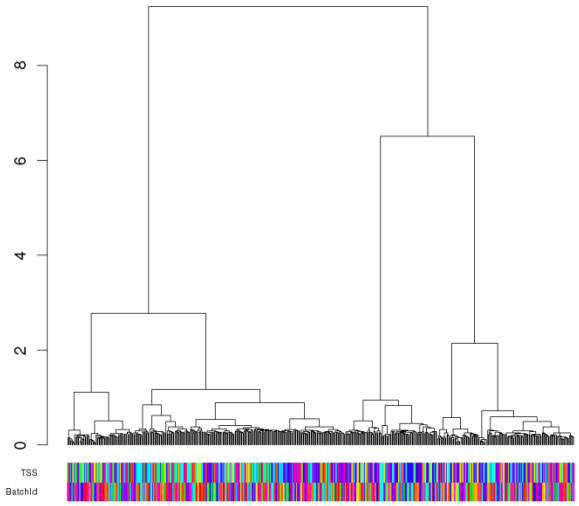


Fig. 3. PCA: First two principal components for miRNA expression from miRNA-seq data, with samples connected by centroids according to TSS.

DNA Methylation (Infinium HM450 microarray)

Figures 4-6 show clustering and PCA plots for the Infinium DNA methylation platform. Principle Components 1 and 3 are shown, because PC 2 showed patient sex based differences (which was expected). None of the batches or tissue source sites stood apart from the others, indicating no serious batch effects were present.



Legends

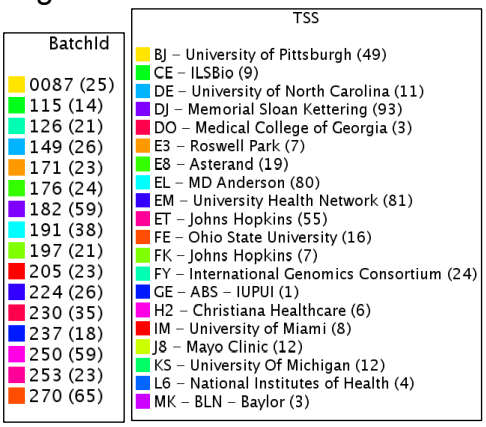


Fig. 4. Hierarchical clustering plot for DNA methylation data.

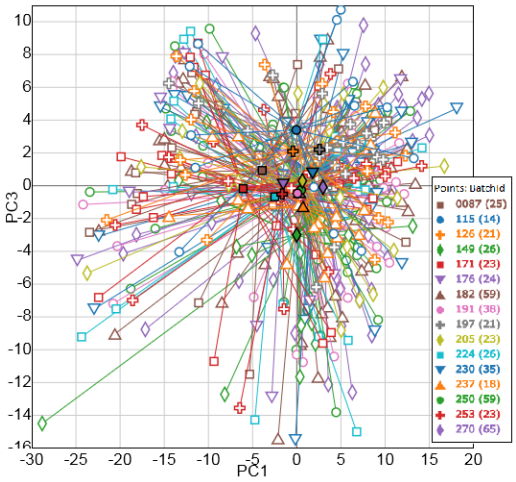


Fig. 5. PCA for DNA methylation, with samples connected by centroids according to batch ID.

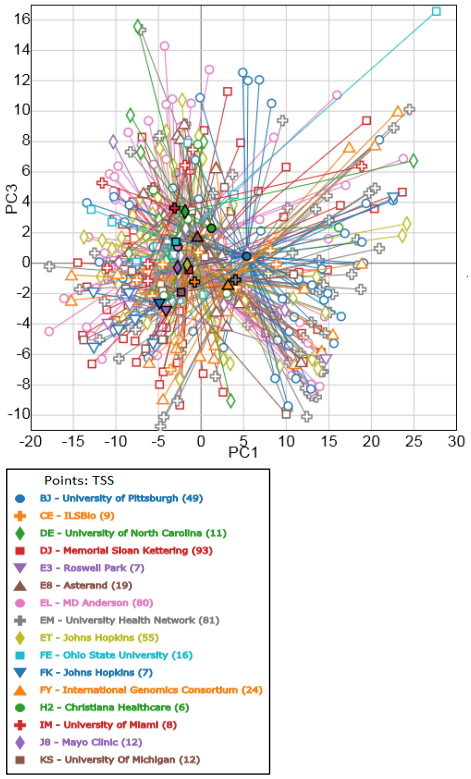
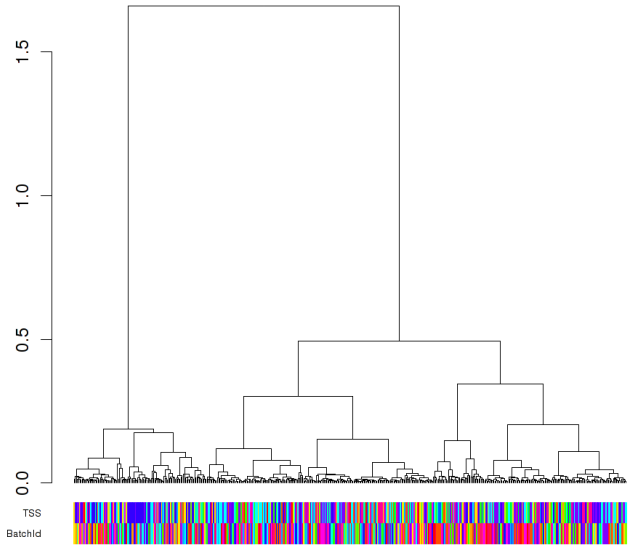


Fig. 6. PCA for DNA methylation, with samples connected by centroids according to TSS.

RNASeqV2 (RNA-Seq Illumina HiSeq)

Figures 7-9 show clustering and PCA plots for the RNA-seq platform. Genes with zero values were removed and the values were $\log_2$-transformed before generating the figures. Once again, no major batch effects were seen.



Legends

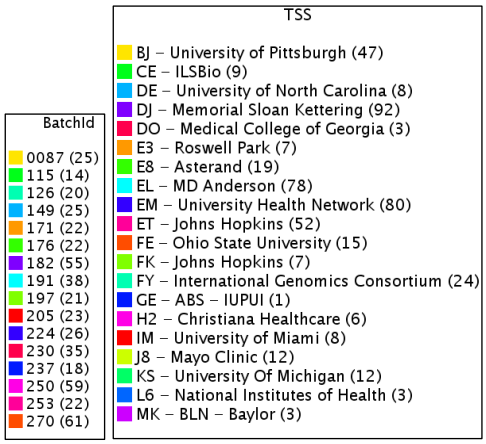


Fig. 7. Hierarchical clustering for mRNA expression from RNA-seq data

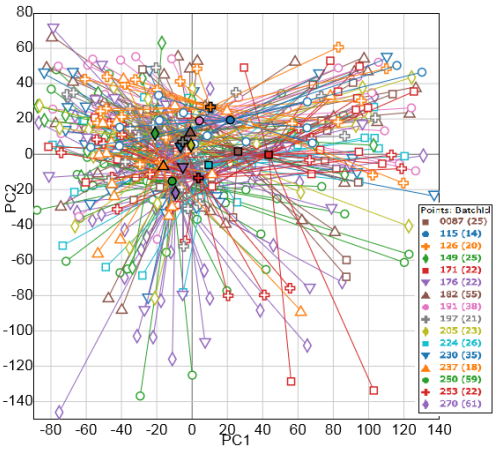


Fig. 8. PCA: First two principal components for RNA-seq, with samples connected by centroids according to batch ID.

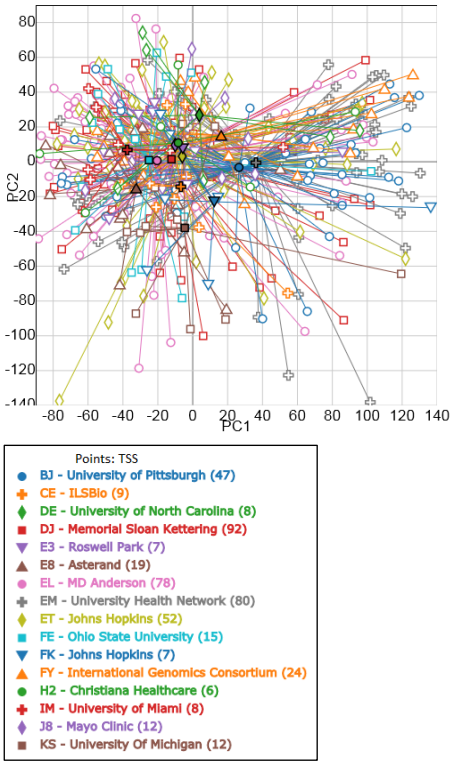
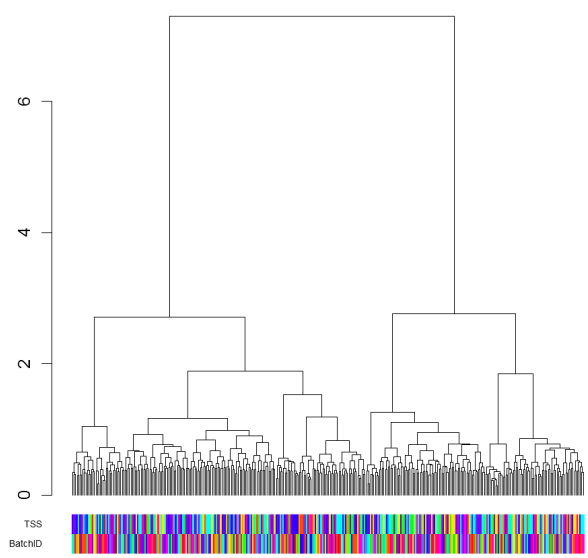


Fig. 9. PCA: First two principal components for RNA-seq, with samples connected by centroids according to TSS.

Protein expression (RPPA)

Figures 10-12 show clustering and PCA plots for protein expression data (Reverse-Phase Protein Array platform). No major batch effects were seen in the hierarchical clustering plot. The PCA plots seem to show some outliers, but that is an artifact of the RPPA platform itself. The PCA plots don't show any major batch effects either.



Legends

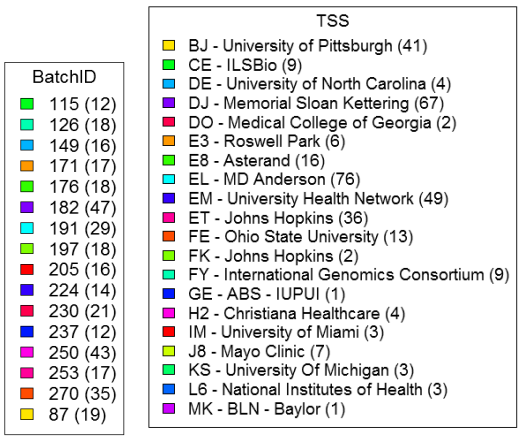


Fig. 10. Hierarchical clustering for RPPA data

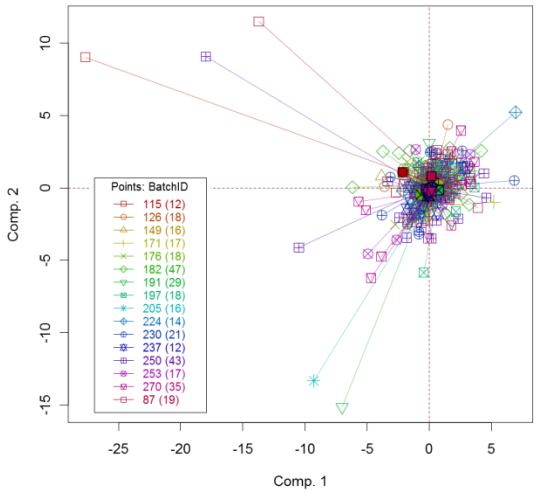


Fig. 11. PCA: First two principal components for RPPA data, with samples connected by centroids according to batch ID.

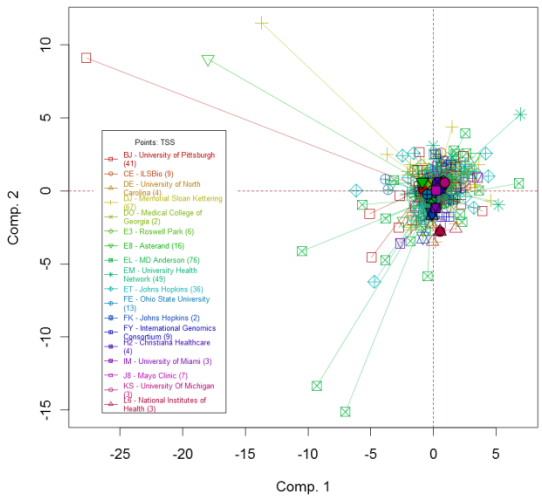


Fig. 12. PCA: First two principal components for RPPA data, with samples connected by centroids according to TSS.

Conclusions

Batch effects were analyzed in four different THCA data sets, miRNA seq, DNA methylation 450k, mRNA seq, and protein expression (RPPA). None of the platforms showed any major batch effects, either by batch ID or by TSS.